D240–D244 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh054

SURFACE: a database of protein surface regions for functional annotation

Fabrizio Ferrè, Gabriele Ausiello, Andreas Zanzoni and Manuela Helmer-Citterich*

Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, 00133 Rome, Italy

Received August 12, 2003; Revised and Accepted September 23, 2003

ABSTRACT

The SURFACE (SUrface Residues and Functions Annotated, Compared and Evaluated, URL http://cbm.bio.uniroma2.it/surface/) database is a repository of annotated and compared protein surface regions. SURFACE contains the results of a large-scale protein annotation and local structural comparison project. A non-redundant set of protein chains is used to build a database of protein surface patches, defined as putative surface functional sites. Each patch is annotated with sequence and structure-derived information about function or interaction abilities. A new procedure for structure comparison is used to perform an all-versus-all patches comparison. Selection of the results obtained with stringent parameters offers a similarity score that can be used to associate different patches and allows reliable annotation by similarity. Annotation exerted through the comparison of regions of protein surface allows the highlighting of similarities that cannot be recognized by other methods of sequence or structure comparison. A graphic representation of the surface patches, functional annotations and the structural superpositions is available through the web interface.

INTRODUCTION

Extracting information about protein functions directly from structure is becoming a crucial task in the structural genomics era (1–3). Structural comparison may lead to the identification of functional relationships even when no clear sequence similarity is detected (4,5). A limitation of this approach is that function is very often encoded in a small number of residues, and cases are known in which proteins sharing similar fold and/or sequence have completely different function (6,7) as well as cases in which a clear functional relationship does not involve sequence/structure similarity (8). In such cases, sequence or structure comparison is likely to be inadequate in describing or identifying protein functions and evolutionary relationships between proteins. These tools, while generally useful in protein classification, may fail in inferring protein functions, being unable to spot local differences. Starting from this background, we have built a relational database to store and spread the results of a large-scale local surface comparison experiment, allowing the scientific community to retrieve non-obvious functional similarities detected between proteins of known 3D structure.

LOCAL SURFACE COMPARISON

We developed a procedure to spot local structural similarities, which is focused on putative functional residues. This method relies on the automatic identification, annotation and structural comparison of functional sites. The first step is the identification of protein surface clefts using the SURFNET algorithm (9), with the demonstrated assumption that there is a clear correspondence between cleft volume and functional involvement (10). Cleft boundaries are explored, identifying those residues that surround the cavity, and that compose the so-called surface 'patch'. For each patch functional information is retrieved from the PROSITE database (11) and from the structure itself, assessing binding abilities from analysis of the bound ligands in the crystal. With the integration of all this information it is possible to obtain a collection of annotated functional sites as surface local patches. This analysis generates a functional sites compendium that can be used to scan a protein structure in order to automatically infer the function(s) of a protein given its structure.

Using a cut-off on volume size to select the biggest clefts, we identify 10 175 surface patches from a non-redundant list of 1924 protein chains whose structure is available. Each patch is composed of an average of 26.5 residues, with a residue distribution that is similar to the residue distribution on the protein surface (a higher frequency of charged and polar residues and a lower frequency of hydrophobic and bulky residues, with respect to the buried residues) (12), although some distinctive features can be detected (i.e. the W frequency in the surface clefts is lower than the W frequency on the overall surface, while the G frequency is higher: follow the Statistics link in the SURFACE home page). We were able to associate at least one functional annotation with 14.4% of these hypothetical functional sites. Using a newly developed structure comparison algorithm (described below) we compare each annotated patch with the whole patches database. Algorithm parameters [such as the root mean square deviation (r.m.s.d.) and minimum similarity of the superposed residues] are set to stringent values, to find only reliable similarities. Moreover, in order to focus the comparison on the putative functional sites, the algorithm is forced to include the annotated residues in the superposition. The similarity

*To whom correspondence should be addressed. Tel: +39 06 72594314; Fax: +39 06 2023500; Email: citterich@uniroma2.it

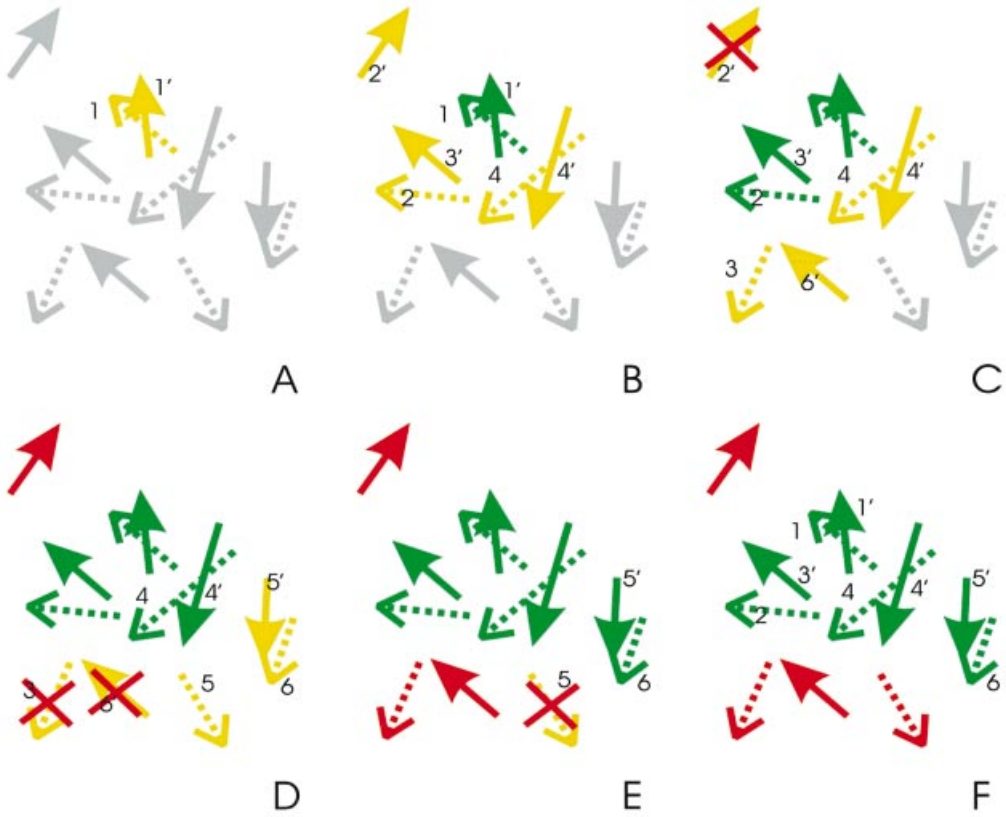

Figure 1. The residues of the two patches P and P′ are represented as straight and dotted arrows, respectively. Each arrow describes the vector joining the alpha carbon and the pseudoatom calculated as the average coordinate of residue side-chain atoms. The colors of the arrows indicate the status of the different residues in the exploration procedure that selects the best matches between the two patches (grey: residue not yet analyzed; green: residue selected for the match; red: residue excluded from the match; yellow: neighbor of the matching residues). A red cross identifies amino acids that have just been discarded. For the sake of clarity, the two patches are always shown (from A to F) in their best superposition. Please note that, in the procedure, the best superposition is calculated for each pair of residues the algorithm explores when trying to extend the match. (**A**) Each possible pair of residues is evaluated. In the example the 1–1′ couple is selected to be the first pair of the match (r.m.s.d. and residue similarity better than a fixed threshold). (**B**) The seed pair is identified and neighboring amino acids are singled out with a distance criterion: residues 2 and 4 of the former patch and 2′, 3′ and 4′ of the latter, and are added to the neighbor list. (**C**) All possible associations between the neighbor residues are tested trying to extend the match. Pairs 2–3′ and 4–4′ are selected and residue 2′ is discarded. The pair 2–3′ is first added to the match. Residues 3 and 6′ are added to the neighbor list. (**D**) Residues 3 and 6′ are discarded, pair 4–4′ is added to the match, residues 5, 5′ and 6 are added to the neighbor list. (**E**) Pair 5′–6 is added to the match, residue 5′ is discarded. (**F**) The final match length is 4 and is composed of residue pairs 1–1′, 2–3′, 4–4′ and 6–5′.

between patches is evaluated by means of the number of superposed residues (the score), and through an evaluation of the match statistical significance based on the score distribution for a given patch (the Z-score).

To test the reliability of the procedure, we verified that for 90% of the annotated patches, at least one patch with the same annotation can be found among the highest-scoring matches (in cases where the annotation being examined is associated with at least two different patches), i.e. the algorithm is able to detect similarity between patches sharing the same annotation. We filtered the huge number of results selecting on the basis of the Z-score: we set a Z-score threshold value calculating, for each different annotation, the average Z-score value of matches between patches sharing the same annotation. Then we fetch those non-annotated patches matching at least two patches with the same annotation and with a Z-score greater than or at least equal to the threshold Z-score for that annotation. We use these conditions to filter the results of the annotated patches-versus-all comparison. A manual analysis of each match has been done, using the literature and information derived from different sources and databases, to determine whether the detected structural similarities can be associated to a functional relationship. For each of the 426 selected matches, a functional relationship between the query and the target patch can be retrieved, validating the procedure and the filtering conditions. The reliability of the procedure in finding meaningful similarities in our non-redundant annotated patches data set has been tested using a set of benchmark cases, in which cryptic similarity between unrelated proteins has been reported, as nucleotide/nucleoside triphosphate binding related to the P-loop (8).

ALGORITHM

A fast and sequence/fold-independent algorithm for local surface comparison has been developed. The algorithm is suitable for large-scale structural comparison given its speed and ability to explore all the combinations of similar/identical

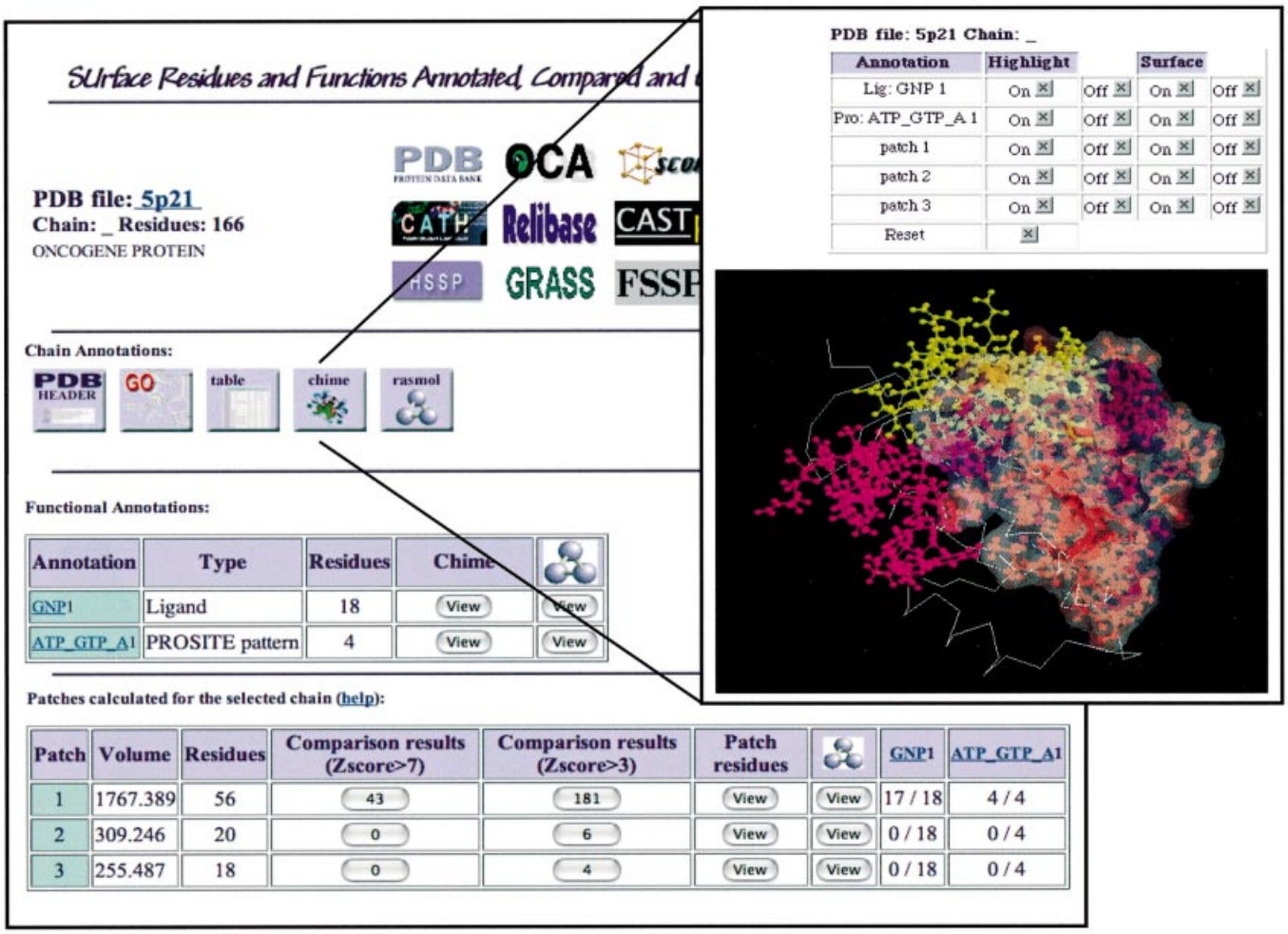

Figure 2. In the left panel, a summary of the information on the selected protein chain is shown. A list of links to other databases is present in the upper part of the page. Four buttons link to the PDB header, GO terms and information about the selected chain annotations (in tabular and graphic format, as shown in the right panel described below). Two tables display SURFACE database data: the first table lists the annotations, the second shows the patches, sorted according to their evaluated volume. Through this last table, the user gains access to the comparison data, only those with Z-score > 7, or the longer list of those with a Z-score > 3. A graphic view of all the annotations associated to the selected chain residues is accessible by clicking on the appropriate button. On the right, the graphic display of the chain annotations can be viewed in a new page with the CHIME and RasMol plug-ins. The single patches, their surfaces and annotated residues can be labeled and displayed in colors.

residues in a sequence-independent way. Two subsets of amino acids are considered to match when their superposition can be associated with a low r.m.s.d. and a good residue similarity according to a chosen substitution matrix (Fig. 1). The first step of the procedure is the reduction of the spatial information: each residue is represented as a pseudo-residue composed of two points: the Cα atom and the geometric center of the side chain atoms (Fig. 1A). The algorithm starts by comparing all possible residue pairs of the query and the target patches. Good seed matches are then selected on the basis of their r.m.s.d. and residue similarity (Fig. 1A). These initial matches are then expanded sequentially by scanning all the remaining residues within 7.5 Å of the seed match (Fig. 1B). At each step (Fig. 1C–E), a new expanded match is accepted or rejected by setting a cut-off value for the r.m.s.d. (typically 0.7 Å) and for the residue similarity (typically 1.2) according to a Dayhoff substitution matrix. The algorithm stops when all possible combinations of subsets have been explored (Fig. 1F).

DATABASE CONTENT AND INTERFACE

Functional annotations and structural comparison results are stored in a freely accessible database; an intuitive interface allows the user to access this information. The database is built with PostgreSQL, an open source object-relational database management system which uses SQL (Structured Query Language) as the query system. The relational structure (not shown) allows easy expansion of the annotation system: new functional annotations can be added without altering the database structure. A collection of Python scripts has been developed to query the database, as well as to create the web interface.

A PDB code can be submitted, or a PDB file can be retrieved through a keyword search. If the selected PDB chain is a representative member in the non-redundant data set, the user can access the chain data, analysis and comparison data. Otherwise the representative member of the redundancy group

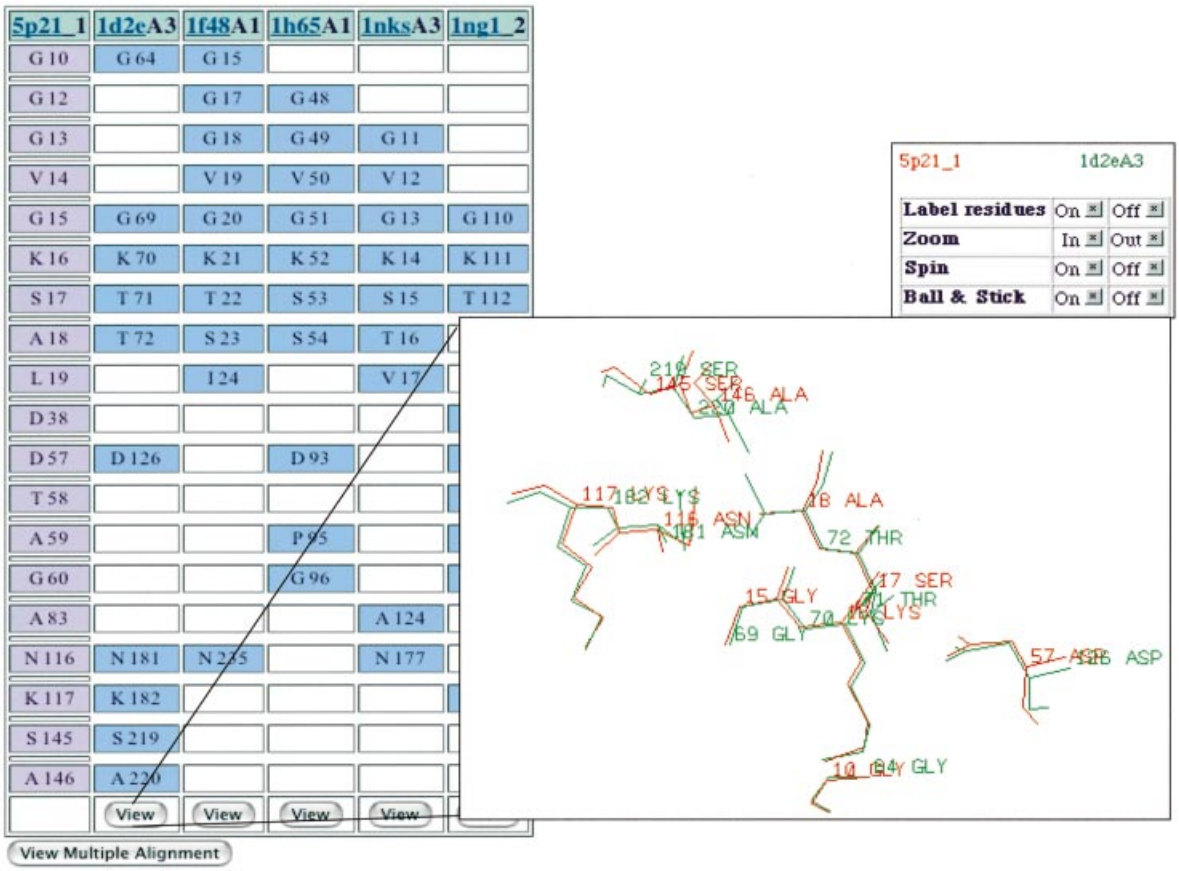

Figure 3. A table of the residues associated in the surface comparison matches and selected in the table with the comparison results. Here the five most significant matches obtained for the 5p21 widest patch are shown. The best match involves the third biggest patch of 1d2e, the mitochondrial EF-Tu protein. The match involves residues that are colinear in the superposed protein chains, but that cannot be aligned (sequence similarity 12.2). By clicking on the appropriate 'view' button, the user can display the superposition of the equivalenced residues.

is proposed. The complete PDB chain data set can be accessed to select a protein. Moreover, the user can access the list of the chains that bind a specific ligand or that match a chosen PROSITE pattern. A form allows the user to submit a protein sequence, retrieving the chains with the highest similarity in the non-redundant list by means of a BLAST search (13). Once a protein chain has been selected, a schematic summary of the annotations and extracted patches is shown (Fig. 2), while a table with the entire information residue per residue can be accessed. Graphical representations of the protein-extracted patches and functional annotations are accessible through the browser plug-in CHIME, based on the program RasMol (14) (Fig. 2), and RasMol.

For each surface patch, the user can retrieve all the structural comparison results with a Z-score > 3 or > 7.0. The comparison results are divided into two blocks: the upper panel shows the patches found structurally similar using the selected patch as bait, sorted by annotation and by Z-score; the lower part shows the patches that fish the query patch, sorted by Z-score. The similarity between patches is scored via the number of the superposed residues (score). Data about the global sequence similarity between the protein chains encompassing the patches is displayed in order to help the user highlighting non-obvious cases (sequence similarity). The user can select one or more matches, and a table, showing the superposed residues, is displayed (Fig. 3). A graphic representation of the detected structural similarities can be visualized using CHIME or RasMol. In Figure 3 a screen snapshot shows the structural similarity between the unrelated proteins human p21 RAS (PDB code: 5p21) and the bovine mitochondrial Ef-Tu protein (PDB code: 1d2e).

CONCLUSIONS AND FUTURE DIRECTIONS

Given the amount of the stored data and the user-friendly web interface, the SURFACE database can be a useful resource for scientific research, providing information about protein functions inferred from different sources and allowing a structural alignment to be obtained easily. This approach has been used to infer the function(s) of a set of uncharacterized proteins whose structure has been solved in structural genomics projects. By adding new categories of functional annotations previously undiscovered similarities can be found. The next database release will include annotations derived from the ELM functional motif database (15) and from the SwissProt database features (16), as well as protein–protein interaction information derived from multimeric complexes in the PDB (17) and from the MINT database on protein–protein interactions (18). The upload of a protein structure, its comparison against the SURFACE database and the retrieval of the similarities detected will be available soon.

ACKNOWLEDGEMENTS

This work was supported by the Telethon multi-centre project GP0101Y01, the EEC project QLRT-2001-02910 and by the AIRC (Associazione Italiana per la Ricerca sul Cancro).

REFERENCES

1. Schmid,M.B. (2002) Structural proteomics: the potential of high-throughput structure determination. *Trends Microbiol.*, **10** (Suppl.), S27–S31.
2. Kinoshita,K., Furui,J. and Nakamura,H. (2001) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics*, **2**, 9–22.
3. Schmitt,S., Kuhn,D. and Klebe,G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.
4. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
5. Holm,L. and Sander,C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins*, **28**, 72–82.
6. Kauvar,L.M. and Villar,H.O. (1998) Deciphering cryptic similarities in protein binding sites. *Curr. Opin. Biotechnol.*, **9**, 390–394.
7. Whisstock,J. and Lesk,A. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.*, in press.
8. Via,A., Ferre,F., Brannetti,B., Valencia,A. and Helmer-Citterich,M. (2000) Three-dimensional view of the surface motif associated with the P-loop structure: *cis* and *trans* cases of convergent evolution. *J. Mol. Biol.*, **303**, 455–465.
9. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 307–308, 323–330.
10. Laskowski,R.A., Luscombe,N.M., Swindells,M.B. and Thornton,J.M. (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438–2452.
11. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
12. Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) Interior and surface of monomeric proteins. *J. Mol. Biol.*, **196**, 641–656.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Sayle,R. and Milner-White,B.J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
15. Puntervoll,P., Linding,R., Gemuend,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M.A., Ausiello,G., Brannetti,B., Costantini,A. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
17. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
18. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.